

Classification challenge on
Alzheimer's Disease
using MRIs and Gene Expression data

May 18, 2022

The problem

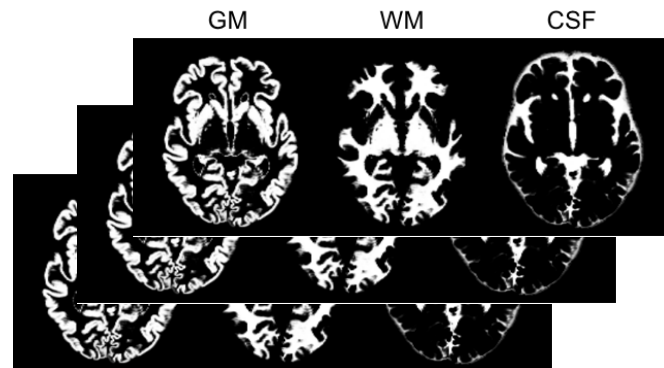
- AD affects about 55M people in the world*
 - need for early diagnosis
- AD (macro-)stages
 - CTL (Controls): no deficit
 - MCI (Mild Cognitive Impairment): few deficits
 - AD (Alzheimer's Disease): dementia

*[WHO2021]

“Hints” from different types of data

- Demographic (age, gender, instruction, ...)
- Clinical evaluation (cognitive tests)
- CSF (Cerebrospinal fluid)
- Medical imaging (**MRIs**, PETs, DTIs, ...)
- Transcriptomics (**gene expression**, ...)

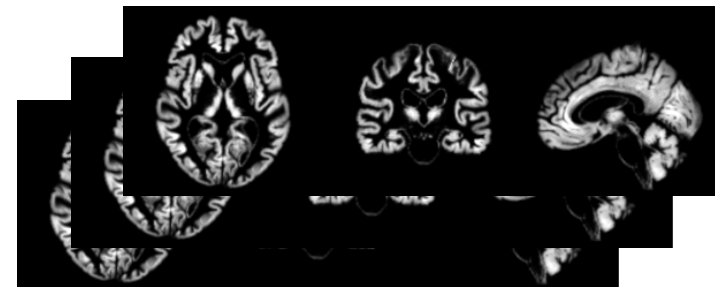
Features from MRIs



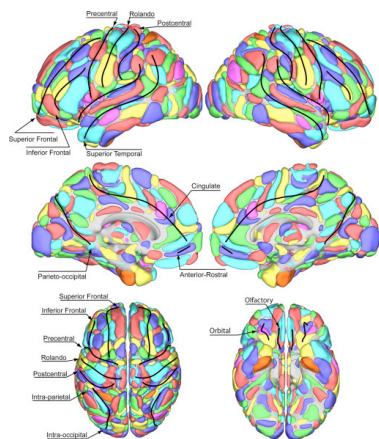
Tissue segmentation, bias correction
and spatial normalization



Inter-subject registration (group template)



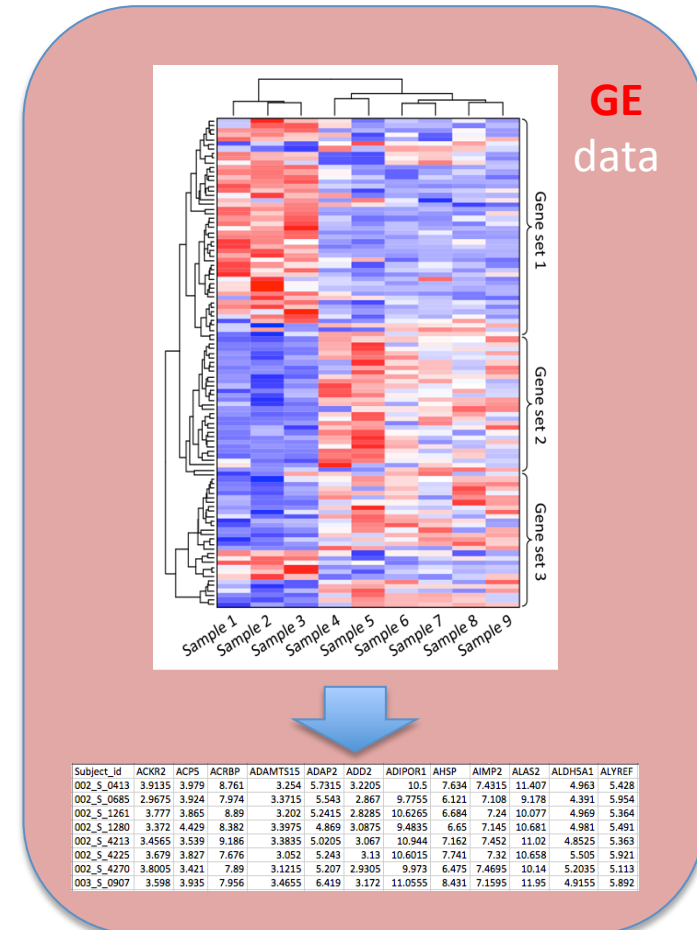
All images mapped to a common space
(MNI), providing a voxel-wise
correspondence across subjects



AICHA Atlas

Average grey
matter densities
obtained from each
anatomical region
defined in an atlas

Combined Data



You will receive

- 3 training datasets for 3 different binary classification problems. These are to be used for training 3 separate classifiers, experimenting different classifiers, feature selection methods, etc. and validating them through cross-validation
- 3 test datasets (one for each binary problem) to test the classifier and provide the obtained predictions

Training data

1) ADCTLtrain.csv: training data for binary classification problem to discriminate AD vs. CN patients

- a) First column: ID of the patient
 - b) Columns from 2 to 430: MRI+GE features
 - c) Last column: Label (patient classification: 'AD' or 'CTL')
- Overall 164 patients: 81 AD and 83CTL

2) ADMCItrain.csv: training data for binary classification problem to discriminate AD vs. MCI patients

- a) First column: ID of the patient
 - b) Columns from 2 to 64: MRI+GE features
 - c) Last column: Label (patient classification: 'AD' or 'MCI')
- Overall 172 patients: 82 AD and 90 MCI

3) MCICTLtrain.csv: training data for binary classification problem to discriminate MCI vs. CTL patients

- a) First column: ID of the patient
 - b) Columns from 2 to 594: MRI+GE features
 - c) Last column: Label (patient classification: 'MCI' or 'CTL')
- Overall 172 patients: 90 MCI and 82 CTL

Example of training dataset

ID	Background	Precentral_L	...	ABCA7	AGTRAP	...	Labels
ADCTL001		AD
ADCTL002		AD
ADCTL003		AD
ADCTL004		AD
ADCTL005		AD
ADCTL006		AD
ADCTL007		AD
ADCTL008		AD
ADCTL009		AD
ADCTL010		AD
...

Test data

1) ADCTLtest.csv: test data for binary classification problem to discriminate AD vs. CTL patients

- a) First column: ID of the patient
- b) Columns from 2 to 430: MRI+GE features

Overall 41 patients

2) ADMCItest.csv: test data for binary classification problem to discriminate AD vs. MCI patients

- a) First column: ID of the patient
- b) Columns from 2 to 64: MRI+GE features

Overall 41 patients

3) MCICTLtest.csv: test data for binary classification problem to discriminate MCI vs. CTL patients

- a) First column: ID of the patient
- b) Columns from 2 to 594: MRI+GE features

Overall 43 patients

Your submission will consist of

For each binary classification problem

1. Two RData files whose name are formatted as:

*StudentRegistrationNumber_FamilyName_**res.Rdata

*StudentRegistrationNumber_FamilyName_**feat.Rdata

where * denotes the classification problem (ADCTL, ADMCI, or MCICL)

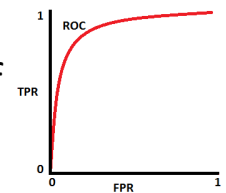
- 1.a) The first file will contain two columns
 - the IDs of the test observation;
 - the predicted labels.
 - 1.b) The second file will contain the column index in the training data files (from 2 to end-1) of the selected features. If features are somehow pre-transformed, describe the transformation in the presentation file (see below).
2. A presentation in PDF with up to 6 pages, in which you describe how you obtained the model. It is NOT mandatory to choose the same classification model for the three different binary problems; just choose the one that leads to the most promising results.
 3. The R macro (script) named *StudentRegistrationNumber_FamilyName_solution.R* used to obtain the results.

The winner is...

- For each binary problem, the results will be ranked according to
 - AUC (e.g., auc R function from pROC) and
 - MCC (e.g., mcc R function from mltools)

$$AUC = \int_0^1 Sens(x) dx, \quad x = 1 - Spec$$

Uses the ROC curve to exhibit the trade-off between the classifier's TP and FP rates



$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Correlation coefficient between observed and predicted binary classifications

Other metrics

$$Acc = \frac{TP + TN}{TP + FN + FP + TN}$$

Percentage of correctly classified samples

$$Spec = \frac{TN}{TN + FP}$$

Percentage of negative samples correctly identified

$$Sens = \frac{TP}{TP + FN}$$

Percentage of positive samples correctly classified (*Recall* or *TPR*)

$$Prec = \frac{TP}{TP + FP}$$

Percentage of positive samples correctly classified, considering the set of all the samples classified as positive

$$F_1 = \frac{2 \cdot Prec \cdot Sens}{Prec + Sens}$$

Compromise between sensitivity and precision

$$BA = \frac{Spec + Sens}{2}$$

Mean of Specificity and Sensitivity

Example results

A) Performance on the training datasets:

	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
AD vs CTL	0.902	0.926	0.880	0.882	0.904	0.969	0.806	0.903
AD vs MCI	0.977	0.976	0.978	0.976	0.976	0.993	0.953	0.977
MCI vs CTL	0.826	0.867	0.780	0.812	0.839	0.884	0.651	0.824

B) Performance on the test datasets using the classifier trained on the training data:

	Acc	Sens	Spec	Prec	F1	AUC	MCC	BA
AD vs CTL	0.829	0.857	0.800	0.818	0.837	0.888	0.659	0.829
AD vs MCI	0.643	0.700	0.591	0.609	0.651	0.761	0.292	0.645
MCI vs CTL	0.628	0.545	0.714	0.667	0.600	0.801	0.263	0.630